Performance Measures to enable Agent-Based Support in Demanding Circumstances

Both, F.; Hoogendoorn, M.; van Lambalgen, R.M.; Oorburg, R.; de Vos, M.

published in
Lecture Notes in Computer Science
2011

DOI (link to publisher)
10.1007/978-3-642-21852-1_67

document version
Publisher's PDF, also known as Version of record

Performance Measures to Enable Agent-Based Support in Demanding Circumstances

Fiemke Both[1], Mark Hoogendoorn[1], Rianne M. van Lambalgen[1], Rogier Oorburg[2], and Michael de Vos[2]

[1] Vrije Universiteit Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
{fboth,mhoogen,rm.van.lambalgen}@few.vu.nl
[2] Defence Materiel Organization, CAMS-Force Vision
P.O. Box 10000, 1780 CA Den Helder, The Netherlands
{r.oorburg,m.de.vos}@forcevision.nl

Abstract. In this paper, an evaluation of measurements that can be used by a personal support agent to measure the quality of human task performance is addressed. Such measurements are important in order for a support agent to give effective and personalized support during the performance of demanding tasks. Hereby, the performance quality measurement is addressed from two perspectives, namely the human's perspective as well as the task perspective. The former represents the idea the human has about the current task performance, whereas the latter measures the actual task performance compared to the goals set for the task at hand. Criteria have been identified to compare the various measurements, and an experiment has been conducted for evaluation. Based on these evaluation results, the most useful measurements are identified to be adopted within personal support agents.

1 Introduction

When humans perform demanding tasks, it is known that their performance can severely degrade over time when their available resources are being exceeded (see e.g. [1]). Such degrading performance is highly undesired, especially in critical domains. Within the research field of augmented cognition, one of the goals is to develop systems that take such limitations of a human's capacity to process information into account and avoid performance degradation by intervening (e.g. [2]). For this purpose, personal assistant agents (e.g. [3], [4]) can be designed where agents interact with sensors in the environment to monitor the human's performance quality and contribute support in case it is needed. Hereby, having information on the performance quality of the human is of essence in order to give appropriate support.

The measurement of how well a human is performing is however not trivial. The quality can be measured from different perspectives, namely the human's perspective (the judgment of performance the human has) as well as the task perspective (depending on the actual task performance). In the field of augmented cognition, both are useful. In order to accept the help of a personal assistant, the human needs to have the idea that the system "understands" the human, hence the agent needs to contain a

D.D. Schmorrow and C.M. Fidopiastis (Eds.): FAC 2011, HCII 2011, LNAI 6780, pp. 578–587, 2011.

model of the human's experience of performance. Furthermore, discrepancies between the human's idea of performance and the actual task performance can also be a basis for an intervention. Of course, the actual task performance is important from the perspective of the eventual outcome of the task.

A variety of measurements that have been proposed in the past as indicators for performance quality can potentially be utilized by an agent applied in a system that is aware of the human state. Indicators for the human's performance are for instance measured using the NASA-TLX [5], or using physiological measurements such as ECG (to measure heart rate). For measurements from the task perspective, agents can use workflow oriented approaches to measure how well the workflow has been followed (see e.g. [6]). In this paper, the measurements are compared to see how suitable they are for usage in a personal assistant agent. Hereby, criteria are identified to score the various measurements, and an experiment has been conducted using a simulation based training environment to evaluate the measurements for their use in agent-based support.

This paper is organized as follows. First, an overview is given of existing performance measurements in Section 2. Thereafter, the criteria for evaluation of measurements are identified in Section 3. Section 4 presents the simulation based training environment used to conduct the experiments, and an evaluation using the data from the experiment is shown in Section 5. Finally, Section 6 is a discussion.

2 Performance Measurements

First, the performance measurements from the human's perspective are described in this section, followed by the measurements from the task perspective.

2.1 Human's Perspective

When looking at performance from a human perspective, the focus is on performance measurements that can be defined by looking at the human. Such measurements can be subjective (e.g. the agent could ask the human to fill in a questionnaire) or psychophysical (the agent could communicate with measurement devices that measure the heart rate). Also, in previous literature human performance is described by looking at the mental effort someone has put in a task. Hockey [7] states that when looking at task performance it is important to take the efficiency of behavior into account. Instead of only looking at a specific task output, it is important to also look at the costs of achieving such an output (i.e. a person's mental effort).

A **subjective measurement** of performance gives information to the agent on how the human is observing the performance. In order to perform these measurements, the subjective scales NASA-Task Load Index (NASA-TLX, [5]) and Subjective Workload Assessment Technique (SWAT, [8]) can be used. Both scales consist of subscales where aspects of mental workload are rated by the human performing a task. In addition, one of the subscales of the NASA-TLX is a performance measure and asks humans to indicate their own performance. The major disadvantage of the subjective performance measurement is that the person performing a task needs to be interrupted by the agent. This can easily be done in an experimental setting, but is not practical in a real world setting.

Physiological measurements provide the personal assistant agent with information about bodily responses to task execution. Examples are EEG (brain activity), Eye Blink Activity and ECG to measure heart rate (HR). HR is known to increase with increasing task demands and decreasing performance ([9]). Concerning Eye Blink Activity, research shows that the time between two successive blinks increases when visual load increases, but decreases when mental (non-visual) load increases ([10]). Both HR and eye blink activity can be very useful as an indicator for task performance for the personal assistant agent. A disadvantage of psychophysiology is that the measurements required can be intrusive and therefore not very desired to use in real world settings (however, less intrusive measurements are also being developed, see e.g. [11]). In addition, when considering HR, other factors should be taken into account. For instance, HR can be influenced by physical exercise, sleep or coffee as well. This should be taken into account by a support agent that uses psychophysical input to reason about a human's state.

2.2 Task Perspective

Some of the approaches described in the previous section are difficult to measure, especially in applications in the real world. Measurements from a task perspective are less intrusive and provide a different type of information about the performance. Depending on the precise reason for which the personal assistant wants to use the task performance for its support actions, one or more of the approaches described below can be used. For example, in a stressful situation it may not be important at all whether the correct procedure is followed, only the outcome matters. Three types of performance measurements from a task perspective are considered here. Section 5.1 gives a detailed description on these measurements applied to the case study.

Effectiveness. The correctness of handling a task based on the set goal, is referred to as *effectiveness*. In this paper, two different perspectives on effectiveness are taken into account. The first is an *absolute* perspective by looking at the outcome regardless of the process leading to that outcome. As the absolute correct outcome is not always available during task execution, it is difficult for an agent to use this information to measure real-time performance. The second perspective is a more realistic perspective on effectiveness, called *realistic* effectiveness. Here, the correctness of a response depends on the workflow that is followed (the correctness of the individual steps that are taken to achieve the outcome).

Productivity. Performance can also be viewed by taking into account the *productivity*. Productivity is often seen as the ratio between the output of a task and the input of a task [12]; the faster the input of a specific task is processed, the more output is generated within a time unit and the higher the productivity. In this paper, two different productivity measurements are taken into account: *average completion time*, and *percentage of cases handled*. These measurements both evaluate the amount of data that is processed within the task.

Efficiency. In addition to productivity and effectiveness, performance from a task-based perspective can be measured by looking into *efficiency*. In this paper, efficiency is defined as the costs of performing a specific task relative to the minimal amount of costs that are necessary to perform the task. Costs are represented by the resources spent on a task, for example money or material.

3 Performance Measurement Evaluation Criteria

In order to compare the measurements for task performance to see how suitable they are for usage in personal assistant agents, a number of criteria have been identified. Hereby, first of all inspiration has been drawn from the work done by [13] in which criteria have been identified to evaluate workload assessment techniques. These can be reused for evaluating performance measurements to be utilized by an agent and are listed below. Note that only the relevant subset of the criteria is taken.

Sensitivity. The sensitivity refers to the capability of the measurement to detect differences in the performance of the human. Some measurements might be relatively coarse grained whereas other can measure on a fine granularity. In this case, two types of sensitivity are involved, namely the sensitivity for the human's perception of performance as well as the sensitivity for the actual task performance. Both are important as argued in the introduction already. As a baseline for actual task performance, the precise performance of the human upon the task at hand is used. Hereby the performance measurement (absolute effectiveness) is directly linked to the goal as provided to the human in the beginning of the task. As a golden standard for the human's perception of performance, the NASA-TLX is used, as this is known to be very reliable for measuring subjective performance. In order to precisely measure how accurate a measurement *m* is for the human's perceived performance and the actual task performance, a linear regression method is used namely *simple linear regression*. The parameters of the linear regression model were a curve of the form $\hat{y} = b_0 + b_1 \cdot x$. Hereby, the x-scale denotes the observed value of measurement *m*, whereas the y-scale indicates the value of the golden standard at the same time point (in this case the precise performance upon the task at hand). $\hat{y}$ denotes the predicted y value for measurement x, and b_0 and b_1 are estimated by means of the *ordinary least square* method and are calculated as follows:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

and:

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \tag{2}$$

In the equation, (x_i, y_i) are pairs of observed measurements (*n* in total). The suitability of the measurement is now defined by taking the average squared error:

$$error = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{3}$$

Intrusiveness. One of the criteria related to the sensing devices themselves is the *intrusiveness* of the sensor to perform measurements. In case the sensors are very intrusive, this might lead to the human feeling uncomfortable, as there is a continuous awareness of everything being measured. The sensors are scored by taking into account how much the human is disturbed during the task itself (e.g. freezing the computer screen to allow the assistant agent to pose a question), and how visible the sensors are. A five point scale is used to score this criterion, ranging from '--' for very intrusive, to '++' for highly non-intrusive ('o' is for neutral).

Reliability. When measurements are performed another important criteria is how robust the measurements are. Some measurements might only be robust when they are performed under laboratory conditions whereas the developed assistant agent might be meant for more demanding conditions. There is often a trade-off between the intrusiveness of sensors, and their robustness. Measuring heart rate using electrodes is more robust compared to measuring it via a watch. However, the latter is less intrusive compared to having electrodes attached to your body. Again, a five point scale is used, whereby '--' stands for not reliable and '++' stands for very reliable.

Implementation requirements. Another criterion includes the requirements of the measurement to be performed, and how difficult it is for an assistant agent to interpret the sensing data. Some data is very easily understandable (e.g. the heart rate), whereas other measurements require the assistant agent to have a more thorough knowledge on how to use the information (e.g. an EEG). In this case, the five point scale ranges from '--' representing heavy implementation requirements to '++' for hardly any requirements.

Task dependence. In the design of personal assistant agent, the goal is often not only to investigate support for a single task, but for multiple tasks to allow for more generic support. Therefore, it is important that the agent does not entirely depend on measurements that highly depend on the characteristics of the tasks being performed. Therefore, the portability of the measurements to other tasks is also included as a criterion. Also here, the same five point scale is used, whereby '--' indicates highly task dependent whereas '++' stands for task independent.

Cost. The last factor is cost. Some sensors are relatively cheap, whereas others can be quite expensive. Again, the five point scale '--' to '++' is used for very high costs to very low costs respectively.

4 Experiment

This section briefly describes the setup of the experiment that has been conducted to evaluate the various measurements. First, the task environment is discussed, followed by the concrete measurements that were performed. Finally, the setup of the experiment is described.

4.1 Simulation-Based Training Environment

The main task that was used in this study consists of identifying incoming contacts on a computer screen and, based on the outcome of identification, deciding to eliminate the contact (by shooting) or allowing it to land (by not shooting). Contacts appear at a random location on the top of the screen and fall down to a random location at the bottom. Shooting is performed by means of a stationary weapon placed on the bottom of the screen. Before a contact can be identified, it has to be perceived. This is done by a mouse click at the contact, which reveals a mathematical equation underneath the contact (e.g. 12*3=36). The identification task is to check the correctness of the mathematical equation (which is less difficult in less demanding situations). A correct equation means that the contact is an ally; an incorrect equation indicates that the contact is an enemy. Identification is done by pressing either the left or right arrow for

respectively an ally or enemy. When a contact is identified a green (for an ally) or a red (for an enemy) circle appears around the contact. The contacts that have been identified as an enemy have to be shot before they land. A missile is shot by executing a mouse click at a specific location; the missile will move from the weapon to that location and explode exactly at the location of the mouse click. When a contact is within a radius of 50 pixels of the exploding missile, it is destroyed. The scenario can in the future easily be extended with a personal assistant agent that measures progress, and takes care of some missiles in cases the human is becoming overloaded. A preliminary study addressing a personal assistant agent for this task environment can be seen in [14], note that the proposed performance measurements in this paper have not been incorporated in the personal assistant agent yet.

4.2 Performance Measurements for the Task

As already stated before, the performance from a *human perspective* was measured with use of a subscale of the NASA-TLX (taken as the golden standard). Each 2.5 minutes participants were asked to rate their performance. In order to conduct the sensitivity evaluation described in Section 3, the participants' ratings were scaled to a number between 0 and 100. For physiological measurements, ECG was measured throughout the entire experiment to calculate the heart rate. Eye blinks were measured using a Tobii x.60 tracker.

For all performance measurements from the *task perspective*, a moving average with a time window of 86 seconds was calculated. To calculate the *absolute effectiveness*, a contact that was handled correctly (e.g. a friend was landed and an enemy was shot) was given an acceptance of 1, a contact that was not handled correctly was given an acceptance of 0. In case of the *realistic effectiveness* acceptance depended on the participants' identification of a contact: an acceptance of 1 was given when a contact identified as friend landed or a contact identified as ally was shot; an acceptance of 0 was given when a contact identified as friend was shot and a contact identified as enemy landed. When a contact was missed, realistic effectiveness was 0. As stated in Section 2, *productivity* was separated in two measurements. First, the average handle time was calculated from the average completion time (time from the time point a contact was instantiated to the time point a contact was handled) and the average reactivity time (time from the time point a contact was instantiated to the time point a contact was perceived:

$$avg\_handle\_time = avg\_completion\_time - avg\_reactivity\_time \quad (4)$$

In addition, the percentage of handled cases was calculated:

$$perc\_handled\_cases = handled\_cases / (handled\_cases + expired\_cases) \quad (5)$$

Finally, the *efficiency* was calculated by dividing the amount of bullets by the amount of handled contacts.

4.3 Participants and Procedure

In this study, 2 female participants and 3 male participants with a mean age of 22.8 took part. All participants already had some experience with the experimental environment.

The experiment consisted of 4 blocks of 20 minutes of the simulation-based training environment. In the first 10 minutes of one block, task demands were low (contacts appear every 10 to 20 seconds) and in the second 10 minutes of one block, task demands were high (contacts appear every 2.25 to 4.5 seconds). In the first and third block, the environment froze after every 2.5 minutes, in the second and fourth block no freezes appeared. The purpose of the freezes was to put the experiment on hold and ask the participants questions about the participants' perceived performance quality. The following sentence was shown: "Gameplay frozen. After this message, a computer version of the NASA-TLX was shown, where participants had to indicate their performance and mental effort. In the future this would be a task that performed by the personal assistant agent.

At the start of the experiment, onscreen instructions were given on the task environment and freezes. The instructions were followed by a practice block of two minutes medium task demands to get familiar with the environment. After practice, participants started with the first block. After each block, the participant was given a three minute break before continuing with the next block.

5 Results

In Table 1 the scores of the various measurements upon the criteria are shown that have been identified to measure the suitability of a measurement for the personal assistant. For calculation of sensitivity, mean values were obtained for each performance measurement and regression analysis was performed. The mean squared error (MSE) (as explained in Section 3) was calculated and averaged over participants. The sensitivity is determined by 1-MSE and scores are presented in Table 1. For the sensitivity with respect to the human's perceived performance, 8 data points were taken from each measurement, 1 for each NASA-TLX measurement in one stage. For the sensitivity with respect to the actual task performance, one data point represented an interval of 20 seconds, the first data points of each part were taken out as no objective data was present yet.

The sensitivity values in Table 1 show that absolute effectiveness (golden standard for task performance) is highly sensitive to the human's perceived performance. The relationship suggests that humans are good in rating their own task performance. Realistic effectiveness is highly sensitive to both task as well as perceived performance. In addition, the completion time is also very sensitive to both types of performances. This could be due to a speed-accuracy trade off: when a case is handled faster, there is more chance of making an error which causes a decrease in performance. The sensitivity scores do not reveal much difference between perceived and actual task performance, except that the measurement eye blink has a relatively high sensitivity for perceived performance compared to task performance. When looking into the data it can be seen that human's perceived performance increases as the time between blinks increases. This effect could be indirectly caused by task demands: as task demands increase, both performance and time between blinks increases.

The rationale for the score on the evaluation criteria apart from the sensitivity criterion is as follows. The NASA-TLX scores negative on intrusiveness as answering

Table 1. Performance Measures Evaluation

Measurement	Task. Sens	Human Sens.	Intru-siveness	Reliability	Implementation requirements	Task Depen-dence	Cost
NASA-TLX	0.977	1.0	--	++	++	++	++
Heart rate	0.900	0.834	o	o	++	++	++
Eyeblink	0.887	0.916	+	--	++	++	++
Absolute effectiveness	1.0	0.957	++	++	o	--	o
Realistic effectiveness	0.976	0.915	++	++	o	--	o
Efficiency	0.936	0.894	++	++	o	--	o
%handled cases	0.926	0.811	++	++	o	--	o
Completion Time	0.959	0.872	++	++	o	--	o

the NASA-TLX questions means that the personal assistant agent would have to interrupt the execution of the current task. The NASA-TLX scores well on reliability, implementation requirements, task dependence, and cost [15]. The heart rate measurement scores neutral on the intrusiveness as well as on the reliability. This is because heart rate can be measured non-intrusive and less reliable (e.g. sensors in clothes), or more reliable and more intrusive (e.g. via ECG using electrodes on the chest). The measurement scores well on implementation requirements, task dependence, and cost. Regarding the eye blinks, the sensor scores well on the intrusiveness, implementation requirements, task dependence, and cost. It does however score relatively bad on reliability, as other environmental aspects can affect the amount of eye blinks (e.g. the amount of sun, tiredness).

Finally, absolute and realistic effectiveness, efficiency and both productivity measurements all score low on task dependence. This is because each time these measurements are used in a different task, a new metric needs to be adopted by the agent. Furthermore, they score mediocre on the implementation requirements as well as cost, as often the software environment in which the task is performed needs to be accessed and possibly extended to allow for a precise measurement to be available for usage by the personal assistant agent. The measurements score high on reliability, because they are highly task dependent. The custom made measurement, that has to be designed for each task, does allow for a reliable representation of performance within the specific task.

6 Discussion

For a personal assistant agent in dynamic circumstances it is useful to have access to different measures of task performance to know the current performance of the

human, allowing for such an agent to give dedicated support. This support could for instance avoid degradation of performance, which is important in the field of augmented cognition. This paper describes several performance measurements that were measured in an experimental setting, all aiming at a different aspect of human performance. The measurements were scored based on a number of criteria and evaluated for their use within a personal assistant agent. The paper shows that especially realistic effectiveness can be perfectly used to substitute both subjective and objective performance as the sensitivity to both measurements is very high. With respect to the psychophysiological measurements, especially eye blink was more predictive of subjective performance compared to objective performance. A possible explanation is that the rating of subjective performance is based upon the responses of the body observed by the human. However, it could also be that both the body and the subjective performance respond to the demands of the task. More research has to be done on the causal nature of this relationship.

The relatively high sensitivity score of all measurements shows that they all can be used to replace either the very intrusive NASA-TLX or the absolute effectiveness that is often not measurable in a real world setting. Depending on the purpose of the support system and the task environment, different approaches can be more or less useful. The advantages of the task-based approaches are the low intrusiveness and high reliability. However, they are very task dependent. In other words, the human does not need to be disturbed at all, but for every new task a new measurement needs to be adopted by the agent. The NASA-TLX questionnaire is also very sensitive, but the very low score for intrusiveness can make it difficult to apply in a real world situation such as an operator working in Air Traffic Control. Here, interruption of the operator can have disastrous consequences.

This research shows that there are several different, very useful performance measurements possible for an agent to use in the example simulation-based training environment. For future research, the idea is to incorporate the most promising performance measurements in a personal assistant agent, and see how well this support agent is able to support the human. Note that a preliminary study concerning this has already been performed (see [14]), however in that setting not all promising measurements have been utilized by the personal assistant agent yet.

References

1. Posner, M.I., Boies, S.J.: Components of attention. Psychological Bulletin. 78, 391–408 (1971)
2. Fuchs, S., Kelly, S., Stanney, K.M., Juhnke, J.S., Dylan, D.: Enhancing mitigation in augmented cognition. Journal of Cognitive Engineering and Decision Making 1(3), 309–326 (2007)
3. Modi, P.J., Veloso, M.M., Smith, S.F., Oh, J.: CMRadar: A Personal Assistant Agent for Calendar Management. In: Bresciani, P., Giorgini, P., Henderson-Sellers, B., Low, G., Winikoff, M. (eds.) AOIS 2004. LNCS (LNAI), vol. 3508, pp. 169–181. Springer, Heidelberg (2005)
4. Myers, K., Berry, P., Blythe, J., Conley, K., Gervasio, M., McGuinness, D.L., Morley, D., Pfeffer, A., Pollack, M., Tambe, M.: An Intelligent Personal Assistant for Task and Time Management. AI Magazine Summer, 47–61 (2007)

5. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Resulsts of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, pp. 139–183. North-Holland, Amsterdam (1988)
6. Muehlen, M.: Workflow-based Process Controlling – Or: What You Can Measure You Can Control. In: Fischer, L. (ed.) Workflow Handbook, pp. 61–77 (2001)
7. Hockey, G.R.J.: Compensatory control in the regulation of human performance under stress and high workload: a cognitive energetical framework. Biological Psychology 45, 73–93 (1997)
8. Reid, G.B., Nygren, T.E.: The Subjective Workload Assessment Technique: a scaling procedure for measuring mental workload. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, pp. 185–218. North Holland, Amsterdam (1988)
9. Both, F., Hoogendoorn, M., Lambalgen, R., van, O.R., Vos, M.: de, Relating Personality and Physiological Measurements to Task Performance Quality. In: Proc. of the 31th Annual Conference of the Cognitive Science Society (CogSci 2009), Austin, TX. Cognitive Science Society (2009)) (to appear)
10. Veltman, J.A., Gaillard, A.W.K.: Physiological workload reactions to increasing levels of task difficulty. Ergonomics 41(5), 656–669 (1998)
11. Pandian, P.S., Mohanavelu, K., Safeer, K.P., Kotresh, T.M., Shakunthala, D.T., Gopal, P., Padaki, V.C.: Smart vest: wearable multi-parameter remote physiological monitoring system. Medical Engineering & Physics 30(4), 466–477 (2008)
12. Coelli, T., Prasada Rao, D.S., O'Donnell, C.J., Battese, G.E.: An introduction to efficiency and productivity analysis. Springer, Heidelberg (2005)
13. Eggemeier, F.T., Wilson, G.F., Kramer, A.F., Damos, D.L.: Workload assessment in multi-task environments. In: Damos, D.L. (ed.) Multi-Task Performance, pp. 206–216. CRC Press, Boca Raton (1991)
14. Bosse, T., Both, F., Duell, R., Hoogendoorn, M., Klein, M.C.A., Lambalgen, R., van Mee, A., van der Oorburg, R., Sharpanskykh, A., Treur, J., de Vos, M.: An Ambient Agent System Assisting Humans in Complex Tasks by Analysis of a Human's State and Performance. In: Proceedings of the Second IEEE International Conference on Intelligent Human Computer Interaction (IHCI 2010). Springer, Heidelberg (2010) (to appear)
15. Rubio, S., Diaz, E., Martin, J., Puente, J.M.: Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. Applied Psychology: an International Review 53(1), 61–86 (2004)